Statistical Science
1991, Vol. 6, No. 4, 363–403

Replication and Meta-Analysis in Parapsychology

Jessica Utts

Abstract. Parapsychology, the laboratory study of psychic phenomena, has had its history interwoven with that of statistics. Many of the controversies in parapsychology have focused on statistical issues, and statistical models have played an integral role in the experimental work. Recently, parapsychologists have been using meta-analysis as a tool for synthesizing large bodies of work. This paper presents an overview of the use of statistics in parapsychology and offers a summary of the meta-analyses that have been conducted. It begins with some anecdotal information about the involvement of statistics and statisticians with the early history of parapsychology. Next, it is argued that most nonstatisticians do not appreciate the connection between power and "successful" replication of experimental effects. Returning to parapsychology, a particular experimental regime is examined by summarizing an extended debate over the interpretation of the results. A new set of experiments designed to resolve the debate is then reviewed. Finally, meta-analyses from several areas of parapsychology are summarized. It is concluded that the overall evidence indicates that there is an anomalous effect in need of an explanation.

Key words and phrases: Effect size, psychic research, statistical controversies, randomness, vote-counting.

1. INTRODUCTION

In a June 1990 Gallup Poll, 49% of the 1236 respondents claimed to believe in extrasensory perception (ESP), and one in four claimed to have had a personal experience involving telepathy (Gallup and Newport, 1991). Other surveys have shown even higher percentages; the University of Chicago's National Opinion Research Center recently surveyed 1473 adults, of which 67% claimed that they had experienced ESP (Greeley, 1987).

Public opinion is a poor arbiter of science, however, and experience is a poor substitute for the scientific method. For more than a century, small numbers of scientists have been conducting laboratory experiments to study phenomena such as telepathy, clairvoyance and precognition, collectively known as "psi" abilities. This paper will examine some of that work, as well as some of the statistical controversies it has generated.

Jessica Utts is Associate Professor, Division of Statistics, University of California at Davis, 469 Kerr Hall, Davis, California 95616.

Parapsychology, as this field is called, has been a source of controversy throughout its history. Strong beliefs tend to be resistant to change even in the face of data, and many people, scientists included, seem to have made up their minds on the question without examining any empirical data at all. A critic of parapsychology recently acknowledged that "The level of the debate during the past 130 years has been an embarrassment for anyone who would like to believe that scholars and scientists adhere to standards of rationality and fair play" (Hyman, 1985a, page 89). While much of the controversy has focused on poor experimental design and potential fraud, there have been attacks and defenses of the statistical methods as well, sometimes calling into question the very foundations of probability and statistical inference.

Most of the criticisms have been leveled by psychologists. For example, a 1988 report of the U.S. National Academy of Sciences concluded that "The committee finds no scientific justification from research conducted over a period of 130 years for the existence of parapsychological phenomena" (Druckman and Swets, 1988, page 22). The chapter on parapsychology was written by a subcommittee

chaired by a psychologist who had published a similar conclusion prior to his appointment to the committee (Hyman, 1985a, page 7). There were no parapsychologists involved with the writing of the report. Resulting accusations of bias (Palmer, Honorton and Utts, 1989) led U.S. Senator Claiborne Pell to request that the Congressional Office of Technology Assessment (OTA) conduct an investigation with a more balanced group. A one-day workshop was held on September 30, 1988, bringing together parapsychologists, critics and experts in some related fields (including the author of this paper). The report concluded that parapsychology needs "a fairer hearing across a broader spectrum of the scientific community, so that emotionality does not impede objective assessment of experimental results" (Office of Technology Assessment, 1989).

It is in the spirit of the OTA report that this article is written. After Section 2, which offers an anecdotal account of the role of statisticians and statistics in parapsychology, the discussion turns to the more general question of replication of experimental results. Section 3 illustrates how replication has been (mis)interpreted by scientists in many fields. Returning to parapsychology in Section 4, a particular experimental regime called the "ganzfeld" is described, and an extended debate about the interpretation of the experimental results is discussed. Section 5 examines a meta-analysis of recent ganzfeld experiments designed to resolve the debate. Finally, Section 6 contains a brief account of meta-analyses that have been conducted in other areas of parapsychology, and conclusions are given in Section 7.

2. STATISTICS AND PARAPSYCHOLOGY

Parapsychology had its beginnings in the investigation of purported mediums and other anecdotal claims in the late 19th century. The Society for Psychical Research was founded in Britain in 1882, and its American counterpart was founded in Boston in 1884. While these organizations and their members were primarily involved with investigating anecdotal material, a few of the early researchers were already conducting "forced-choice" experiments such as card-guessing. (Forced-choice experiments are like multiple choice tests; on each trial the subject must guess from a small, known set of possibilities.) Notable among these was Nobel Laureate Charles Richet, who is generally credited with being the first to recognize that probability theory could be applied to card-guessing experiments (Rhine, 1977, page 26; Richet, 1884).

F. Y. Edgeworth, partly in response to what he considered to be incorrect analyses of these experiments, offered one of the earliest treatises on the statistical evaluation of forced-choice experiments in two articles published in the *Proceedings of the Society for Psychical Research* (Edgeworth, 1885, 1886). Unfortunately, as noted by Mauskopf and McVaugh (1979) in their historical account of the period, Edgeworth's papers were "perhaps too difficult for their immediate audience" (page 105).

Edgeworth began his analysis by using Bayes' theorem to derive the formula for the posterior probability that chance was operating, given the data. He then continued with an argument "savouring more of Bernoulli than Bayes" in which "it is consonant, I submit, to experience, to put 1/2 both for α and β," that is, for both the prior probability that chance alone was operating, and the prior probability that "there should have been some additional agency." He then reasoned (using a Taylor series expansion of the posterior probability formula) that if there were a large probability of observing the data given that some additional agency was at work, and a small objective probability of the data under chance, then the latter (binomial) probability "may be taken as a rough measure of the sought *a posteriori* probability in favour of mere chance" (page 195). Edgeworth concluded his article by applying his method to some data published previously in the same journal. He found the probability against chance to be 0.99996, which he said "may fairly be regarded as physical certainty" (page 199). He concluded:

> Such is the evidence which the calculus of probabilities affords as to the existence of an agency other than mere chance. The calculus is silent as to the nature of that agency—whether it is more likely to be vulgar illusion or extraordinary law. That is a question to be decided, not by formulae and figures, but by general philosophy and common sense [page 199].

Both the statistical arguments and the experimental controls in these early experiments were somewhat loose. For example, Edgeworth treated as binomial an experiment in which one person chose a string of eight letters and another attempted to guess the string. Since it has long been understood that people are poor random number (or letter) generators, there is no statistical basis for analyzing such an experiment. Nonetheless, Edgeworth and his contemporaries set the stage for the use of controlled experiments with statistical evaluation in laboratory parapsychology. An interesting historical account of Edgeworth's involvement and the role telepathy experiments played in the early history of randomization and experimental design is provided by Hacking (1988).

One of the first American researchers to use statistical methods in parapsychology was John Edgar Coover, who was the Thomas Welton Stanford Psychical Research Fellow in the Psychology Department at Stanford University from 1912 to 1937 (Dommeyer, 1975). In 1917, Coover published a large volume summarizing his work (Coover, 1917). Coover believed that his results were consistent with chance, but others have argued that Coover's definition of significance was too strict (Dommeyer, 1975). For example, in one evaluation of his telepathy experiments, Coover found a two-tailed p-value of 0.0062. He concluded, "Since this value, then, lies within the field of chance deviation, although the probability of its occurrence by chance is fairly low, it cannot be accepted as a decisive indication of some cause beyond chance which operated in favor of success in guessing" (Coover, 1917, page 82). On the next page, he made it explicit that he would require a p-value of 0.0000221 to declare that something other than chance was operating.

It was during the summer of 1930, with the card-guessing experiments of J. B. Rhine at Duke University, that parapsychology began to take hold as a laboratory science. Rhine's laboratory still exists under the name of the Foundation for Research on the Nature of Man, housed at the edge of the Duke University campus.

It wasn't long after Rhine published his first book, *Extrasensory Perception* in 1934, that the attacks on his methodology began. Since his claims were wholly based on statistical analyses of his experiments, the statistical methods were closely scrutinized by critics anxious to find a conventional explanation for Rhine's positive results.

The most persistent critic was a psychologist from McGill University named Chester Kellogg (Mauskopf and McVaugh, 1979). Kellogg's main argument was that Rhine was using the binomial distribution (and normal approximation) on a series of trials that were not independent. The experiments in question consisted of having a subject guess the order of a deck of 25 cards, with five each of five symbols, so technically Kellogg was correct.

By 1937, several mathematicians and statisticians had come to Rhine's aid. Mauskopf and McVaugh (1979) speculated that since statistics was itself a young discipline, "a number of statisticians were equally outraged by Kellogg, whose arguments they saw as discrediting *their* profession" (page 258). The major technical work, which acknowledged that Kellogg's criticisms were accurate but did little to change the significance of the results, was conducted by Charles Stuart and Joseph A. Greenwood and published in the first volume of the *Journal of Parapsychology* (Stuart and Greenwood, 1937). Stuart, who had been an undergraduate in mathematics at Duke, was one of Rhine's early subjects and continued to work with him as a researcher until Stuart's death in 1947. Greenwood was a Duke mathematician, who apparently converted to a statistician at the urging of Rhine.

Another prominent figure who was distressed with Kellogg's attack was E. V. Huntington, a mathematician at Harvard. After corresponding with Rhine, Huntington decided that, rather than further confuse the public with a technical reply to Kellogg's arguments, a simple statement should be made to the effect that the mathematical issues in Rhine's work had been resolved. Huntington must have successfully convinced his former student, Burton Camp of Wesleyan, that this was a wise approach. Camp was the 1937 President of IMS. When the annual meetings were held in December of 1937 (jointly with AMS and AAAS), Camp released a statement to the press that read:

> Dr. Rhine's investigations have two aspects: experimental and statistical. On the experimental side mathematicians, of course, have nothing to say. On the statistical side, however, recent mathematical work has established the fact that, assuming that the experiments have been properly performed, the statistical analysis is essentially valid. If the Rhine investigation is to be fairly attacked, it must be on other than mathematical grounds [Camp, 1937].

One statistician who did emerge as a critic was William Feller. In a talk at the Duke Mathematical Seminar on April 24, 1940, Feller raised three criticisms to Rhine's work (Feller, 1940). They had been raised before by others (and continue to be raised even today). The first was that inadequate shuffling of the cards resulted in additional information from one series to the next. The second was what is now known as the "file-drawer effect," namely, that if one combines the results of published studies only, there is sure to be a bias in favor of successful studies. The third was that the results were enhanced by the use of optional stopping, that is, by not specifying the number of trials in advance. All three of these criticisms were addressed in a rejoinder by Greenwood and Stuart (1940), but Feller was never convinced. Even in its third edition published in 1968, his book *An Introduction to Probability Theory and Its Applications* still contains his conclusion about Greenwood and Stuart: "Both their arithmetic and their experiments have a distinct tinge of the supernatural" (Feller, 1968, page 407). In his discussion of Feller's position, Diaconis (1978) remarked, "I believe

Feller was confused . . . he seemed to have decided the opposition was wrong and that was that."

Several statisticians have contributed to the literature in parapsychology to greater or lesser degrees. T. N. E. Greville developed applicable statistical methods for many of the experiments in parapsychology and was Statistical Editor of the *Journal of Parapsychology* (with J. A. Greenwood) from its start in 1937 through Volume 31 in 1967; Fisher (1924, 1929) addressed some specific problems in card-guessing experiments; Wilks (1965a, b) described various statistical methods for parapsychology; Lindley (1957) presented a Bayesian analysis of some parapsychology data; and Diaconis (1978) pointed out some problems with certain experiments and presented a method for analyzing experiments when feedback is given.

Occasionally, attacks on parapsychology have taken the form of attacks on statistical inference in general, at least as it is applied to real data. Spencer-Brown (1957) attempted to show that true randomness is impossible, at least in finite sequences, and that this could be the explanation for the results in parapsychology. That argument reemerged in a recent debate on the role of randomness in parapsychology, initiated by psychologist J. Barnard Gilmore (Gilmore, 1989, 1990; Utts, 1989; Palmer, 1989, 1990). Gilmore stated that "The agnostic statistician, advising on research in psi, should take account of the possible inappropriateness of classical inferential statistics" (1989, page 338). In his second paper, Gilmore reviewed several non-psi studies showing purportedly random systems that do not behave as they should under randomness (e.g., Iversen, Longcor, Mosteller, Gilbert and Youtz, 1971; Spencer-Brown, 1957). Gilmore concluded that "Anomalous data . . . should not be found nearly so often if classical statistics offers a valid model of reality" (1990, page 54), thus rejecting the use of classical statistical inference for real-world applications in general.

3. REPLICATION

Implicit and explicit in the literature on parapsychology is the assumption that, in order to truly establish itself, the field needs to find a repeatable experiment. For example, Diaconis (1978) started the summary of his article in *Science* with the words "In search of repeatable ESP experiments, modern investigators . . ." (page 131). On October 28–29, 1983, the 32nd International Conference of the Parapsychology Foundation was held in San Antonio, Texas, to address "The Repeatability Problem in Parapsychology." The Conference Proceedings (Shapin and Coly, 1985) reflect the diverse views among parapsychologists on the nature of the problem. Honorton (1985a) and Rao (1985), for example, both argued that strict replication is uncommon in *most* branches of science and that parapsychology should not be singled out as unique in this regard. Other authors expressed disappointment in the lack of a single repeatable experiment in parapsychology, with titles such as "Unrepeatability: Parapsychology's Only Finding" (Blackmore, 1985), and "Research Strategies for Dealing with Unstable Phenomena" (Beloff, 1985).

It has never been clear, however, just exactly what would constitute acceptable evidence of a repeatable experiment. In the early days of investigation, the major critics "insisted that it would be sufficient for Rhine and Soal to convince them of ESP if a parapsychologist could perform successfully a single 'fraud-proof' experiment" (Hyman, 1985a, page 71). However, as soon as well-designed experiments showing statistical significance emerged, the critics realized that a single experiment could be statistically significant just by chance. British psychologist C. E. M. Hansel quantified the new expectation, that the experiment should be repeated a few times, as follows:

> If a result is significant at the .01 level and this result is not due to chance but to information reaching the subject, it may be expected that by making two further sets of trials the antichance odds of one hundred to one will be increased to around a million to one, thus enabling the effects of ESP—or whatever is responsible for the original result—to manifest itself to such an extent that there will be little doubt that the result is not due to chance [Hansel, 1980, page 298].

In other words, three consecutive experiments at $p \leq 0.01$ would convince Hansel that something other than chance was at work.

This argument implies that if a particular experiment produces a statistically significant result, but subsequent replications fail to attain significance, then the original result was probably due to chance, or at least remains unconvincing. The problem with this line of reasoning is that there is no consideration given to sample size or power. Only an experiment with extremely high power should be expected to be "successful" three times in succession.

It is perhaps a failure of the way statistics is taught that many scientists do not understand the importance of power in defining successful replication. To illustrate this point, psychologists Tversky and Kahnemann (1982) distributed a questionnaire

to their colleagues at a professional meeting, with the question:

> An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value, $t = 2.46$. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of t with the same number of subjects. The direction was the same in both sets of data. You are reviewing the literature. What is the highest value of t in the second set of data that you would describe as a failure to replicate? [1982, page 28].

In reporting their results, Tversky and Kahnemann stated:

> The majority of our respondents regarded $t = 1.70$ as a failure to replicate. If the data of two such studies ($t = 2.46$ and $t = 1.70$) are pooled, the value of t for the combined data is about 3.00 (assuming equal variances). Thus, we are faced with a paradoxical state of affairs, in which the same data that would increase our confidence in the finding when viewed as part of the original study, shake our confidence when viewed as an independent study [1982, page 28].

At a recent presentation to the History and Philosophy of Science Seminar at the University of California at Davis, I asked the following question. Two scientists, Professors A and B, each have a theory they would like to demonstrate. Each plans to run a fixed number of Bernoulli trials and then test H_0: $p = 0.25$ versus H_a: $p > 0.25$. Professor A has access to large numbers of students each semester to use as subjects. In his first experiment, he runs 100 subjects, and there are 33 successes ($p = 0.04$, one-tailed). Knowing the importance of replication, Professor A runs an additional 100 subjects as a second experiment. He finds 36 successes ($p = 0.009$, one-tailed).

Professor B only teaches small classes. Each quarter, she runs an experiment on her students to test her theory. She carries out ten studies this way, with the results in Table 1.

I asked the audience by a show of hands to indicate whether or not they felt the scientists had successfully demonstrated their theories. Professor A's theory received overwhelming support, with approximately 20 votes, while Professor B's theory received only one vote.

If you aggregate the results of the experiments for each professor, you will notice that each conducted 200 trials, and Professor B actually demonstrated a *higher* level of success than Professor A, with 71 as opposed to 69 successful trials. The one-tailed p-values for the combined trials are 0.0017 for Professor A and 0.0006 for Professor B.

To address the question of replication more explicitly, I also posed the following scenario. In December of 1987, it was decided to prematurely terminate a study on the effects of aspirin in reducing heart attacks because the data were so convincing (see, e.g., Greenhouse and Greenhouse, 1988; Rosenthal, 1990a). The physician-subjects had been randomly assigned to take aspirin or a placebo. There were 104 heart attacks among the 11,037 subjects in the aspirin group, and 189 heart attacks among the 11,034 subjects in the placebo group (chi-square = 25.01, $p < 0.00001$).

After showing the results of that study, I presented the audience with two hypothetical experiments conducted to try to replicate the original result, with outcomes in Table 2.

I asked the audience to indicate which one they thought was a more successful replication. The audience chose the second one, as would most journal editors, because of the "significant p-value." In fact, the *first* replication has almost exactly the same proportion of heart attacks in the two groups as the original study and is thus a very close replication of that result. The second replication has

TABLE 1
Attempted replciations for professor B

n	Number of successes	One-tailed p-value
10	4	0.22
15	6	0.15
17	6	0.23
25	8	0.17
30	10	0.20
40	13	0.18
18	7	0.14
10	5	0.08
15	5	0.31
20	7	0.21

TABLE 2
Hypothetical replications of the aspirin / heart attack study

	Replication #1 Heart attack		Replication #2 Heart attack	
	Yes	No	Yes	No
Aspirin	11	1156	20	2314
Placebo	19	1090	48	2170
Chi-square	2.596, $p = 0.11$		13.206, $p = 0.0003$	

very *different* proportions, and in fact the relative risk from the second study is not even contained in a 95% confidence interval for relative risk from the original study. The *magnitude* of the effect has been much more closely matched by the "nonsignificant" replication.

Fortunately, psychologists are beginning to notice that replication is not as straightforward as they were originally led to believe. A special issue of the *Journal of Social Behavior and Personality* was entirely devoted to the question of replication (Neuliep, 1990). In one of the articles, Rosenthal cautioned his colleagues: "Given the levels of statistical power at which we normally operate, we have no right to expect the proportion of significant results that we typically do expect, even if in nature there is a very real and very important effect" (Rosenthal, 1990b, page 16).

Jacob Cohen, in his insightful article titled "Things I Have Learned (So Far)," identified another misconception common among social scientists: "Despite widespread misconceptions to the contrary, the rejection of a given null hypothesis gives us no basis for estimating the probability that a replication of the research will again result in rejecting that null hypothesis" (Cohen, 1990, page 1307).

Cohen and Rosenthal both advocate the use of effect sizes as opposed to significance levels when defining the strength of an experimental effect. In general, effect sizes measure the amount by which the data deviate from the null hypothesis in terms of standardized units. For instance, the effect size for a two-sample t-test is usually defined to be the difference in the two means, divided by the standard deviation for the control group. This measure can be compared across studies without the dependence on sample size inherent in significance levels. (Of course there will still be variability in the sample effect sizes, decreasing as a function of sample size.) Comparison of effect sizes across studies is one of the major components of meta-analysis.

Similar arguments have recently been made in the medical literature. For example, Gardner and Altman (1986) stated that the use of p-values "to define two alternative outcomes—significant and not significant—is not helpful and encourages lazy thinking" (page 746). They advocated the use of confidence intervals instead.

As discussed in the next section, the arguments used to conclude that parapsychology has failed to demonstrate a replicable effect hinge on these misconceptions of replication and failure to examine power. A more appropriate analysis would compare the effect sizes for similar experiments across experimenters and across time to see if there have been consistent effects of the same magnitude. Rosenthal also advocates this view of replication:

> The traditional view of replication focuses on significance level as the relevant summary statistic of a study and evaluates the success of a replication in a dichotomous fashion. The newer, more useful view of replication focuses on effect size as the more important summary statistic of a study and evaluates the success of a replication not in a dichotomous but in a continuous fashion [Rosenthal, 1990b, page 28].

The dichotomous view of replication has been used throughout the history of parapsychology, by both parapsychologists and critics (Utts, 1988). For example, the National Academy of Sciences report critically evaluated "significant" experiments, but entirely ignored "nonsignificant" experiments.

In the next three sections, we will examine some of the results in parapsychology using the broader, more appropriate definition of replication. In doing so, we will show that the results are far more interesting than the critics would have us believe.

4. THE GANZFELD DEBATE IN PARAPSYCHOLOGY

An extensive debate took place in the mid-1980s between a parapsychologist and critic, questioning whether or not a particular body of parapsychological data had demonstrated psi abilities. The experiments in question were all conducted using the ganzfeld setting (described below). Several authors were invited to write commentaries on the debate. As a result, this data base has been more thoroughly analyzed by both critics and proponents than any other and provides a good source for studying replication in parapsychology.

The debate concluded with a detailed series of recommendations for further experiments, and left open the question of whether or not psi abilities had been demonstrated. A new series of experiments that followed the recommendations were conducted over the next few years. The results of the new experiments will be presented in Section 5.

4.1 Free-Response Experiments

Recent experiments in parapsychology tend to use more complex target material than the cards and dice used in the early investigations, partially to alleviate boredom on the part of the subjects and partially because they are thought to "more nearly resemble the conditions of spontaneous psi occurrences" (Burdick and Kelly, 1977, page 109). These experiments fall under the general heading of "free-response" experiments, because the subject is asked to give a verbal or written description of the

target, rather than being forced to make a choice from a small discrete set of possibilities. Various types of target material have been used, including pictures, short segments of movies on video tapes, actual locations and small objects.

Despite the more complex target material, the statistical methods used to analyze these experiments are similar to those for forced-choice experiments. A typical experiment proceeds as follows. Before conducting any trials, a large pool of potential targets is assembled, usually in packets of four. Similarity of targets within a packet is kept to a minimum, for reasons made clear below. At the start of an experimental session, after the subject is sequestered in an isolated room, a target is selected at random from the pool. A sender is placed in another room with the target. The subject is asked to provide a verbal or written description of what he or she thinks is in the target, knowing only that it is a photograph, an object, etc.

After the subject's description has been recorded and secured against the potential for later alteration, a judge (who may or may not be the subject) is given a copy of the subject's description and the four possible targets that were in the packet with the correct target. A properly conducted experiment either uses video tapes or has two identical sets of target material and uses the duplicate set for this part of the process, to ensure that clues such as fingerprints don't give away the answer. Based on the subject's description, and of course on a blind basis, the judge is asked to either rank the four choices from most to least likely to have been the target, or to select the one from the four that seems to best match the subject's description. If ranks are used, the statistical analysis proceeds by summing the ranks over a series of trials and comparing the sum to what would be expected by chance. If the selection method is used, a "direct hit" occurs if the correct target is chosen, and the number of direct hits over a series of trials is compared to the number expected in a binomial experiment with $p = 0.25$.

Note that the subjects' *responses* cannot be considered to be "random" in any sense, so probability assessments are based on the random selection of the target and decoys. In a correctly designed experiment, the probability of a direct hit by chance is 0.25 on each trial, regardless of the response, and the trials are independent. These and other issues related to analyzing free-response experiments are discussed by Utts (1991).

4.2 The Psi Ganzfeld Experiments

The ganzfeld procedure is a particular kind of free-response experiment utilizing a perceptual isolation technique originally developed by Gestalt psychologists for other purposes. Evidence from spontaneous case studies and experimental work had led parapsychologists to a model proposing that psychic functioning may be masked by sensory input and by inattention to internal states (Honorton, 1977). The ganzfeld procedure was specifically designed to test whether or not reduction of external "noise" would enhance psi performance.

In these experiments, the subject is placed in a comfortable reclining chair in an acoustically shielded room. To create a mild form of sensory deprivation, the subject wears headphones through which white noise is played, and stares into a constant field of red light. This is achieved by taping halved translucent ping-pong balls over the eyes and then illuminating the room with red light. In the psi ganzfeld experiments, the subject speaks into a microphone and attempts to describe the target material being observed by the sender in a distant room.

At the 1982 Annual Meeting of the Parapsychological Association, a debate took place over the degree to which the results of the psi ganzfeld experiments constituted evidence of psi abilities. Psychologist and critic Ray Hyman and parapsychologist Charles Honorton each analyzed the results of all known psi ganzfeld experiments to date, and they reached strikingly different conclusions (Honorton, 1985b; Hyman, 1985b). The debate continued with the publication of their arguments in separate articles in the March 1985 issue of the *Journal of Parapsychology*. Finally, in the December 1986 issue of the *Journal of Parapsychology*, Hyman and Honorton (1986) wrote a joint article in which they highlighted their agreements and disagreements and outlined detailed criteria for future experiments. That same issue contained commentaries on the debate by 10 other authors.

The data base analyzed by Hyman and Honorton (1986) consisted of results taken from 34 reports written by a total of 47 authors. Honorton counted 42 separate experiments described in the reports, of which 28 reported enough information to determine the number of direct hits achieved. Twenty three of the studies (55%) were classified by Honorton as having achieved statistical significance at 0.05.

4.3 The Vote-Counting Debate

Vote-counting is the term commonly used for the technique of drawing inferences about an experimental effect by counting the number of significant versus nonsignificant studies of the effect. Hedges and Olkin (1985) give a detailed analysis of the inadequacy of this method, showing that it is more and more likely to make the wrong decision as the

number of studies increases. While Hyman acknowledged that "vote-counting raises many problems" (Hyman, 1985b, page 8), he nonetheless spent half of his critique of the ganzfeld studies showing why Honorton's count of 55% was wrong.

Hyman's first complaint was that several of the studies contained multiple conditions, each of which should be considered as a separate study. Using this definition he counted 80 studies (thus further reducing the sample sizes of the individual studies), of which 25 (31%) were "successful." Honorton's response to this was to invite readers to examine the studies and decide for themselves if the varying conditions constituted separate experiments.

Hyman next postulated that there was selection bias, so that significant studies were more likely to be reported. He raised some important issues about how pilot studies may be terminated and not reported if they don't show significant results, or may at least be subject to optional stopping, allowing the experimenter to determine the number of trials. He also presented a chi-square analysis that "suggests a tendency to report studies with a small sample only if they have significant results" (Hyman, 1985b, page 14), but I have questioned his analysis elsewhere (Utts, 1986, page 397).

Honorton refuted Hyman's argument with four rejoinders (Honorton, 1985b, page 66). In addition to reinterpreting Hyman's chi-square analysis, Honorton pointed out that the Parapsychological Association has an official policy encouraging the publication of nonsignificant results in its journals and proceedings, that a large number of reported ganzfeld studies did not achieve statistical significance and that there would have to be 15 studies in the "file-drawer" for every one reported to cancel out the observed significant results.

The remainder of Hyman's vote-counting analysis consisted of showing that the effective error rate for each study was actually much higher than the nominal 5%. For example, each study could have been analyzed using the direct hit measure, the sum of ranks measure or one of two other measures used for free-response analyses. Hyman carried out a simulation study that showed the true error rate would be 0.22 if "significance" was defined by requiring at least one of these four measures to achieve the 0.05 level. He suggested several other ways in which multiple testing could occur and concluded that the effective error rate in each experiment was not the nominal 0.05, but rather was probably close to the 31% he had determined to be the actual success rate in his vote-count.

Honorton acknowledged that there was a multiple testing problem, but he had a two-fold response. First, he applied a Bonferroni correction and found that the number of significant studies (using his definition of a study) only dropped from 55% to 45%. Next, he proposed that a uniform index of success be applied to all studies. He used the number of direct hits, since it was by far the most commonly reported measure and was the measure used in the first published psi ganzfeld study. He then conducted a detailed analysis of the 28 studies reporting direct hits and found that 43% were significant at 0.05 on that measure alone. Further, he showed that significant effects were reported by six of the 10 independent investigators and thus were not due to just one or two investigators or laboratories. He also noted that success rates were very similar for reports published in refereed journals and those published in unrefereed monographs and abstracts.

While Hyman's arguments identified issues such as selective reporting and optional stopping that should be considered in any meta-analysis, the dependence of significance levels on sample size makes the vote-counting technique almost useless for assessing the magnitude of the effect. Consider, for example, the 24 studies where the direct hit measure was reported and the chance probability of a direct hit was 0.25, the most common type of study in the data base. (There were four direct hit studies with other chance probabilities and 14 that did not report direct hits.) Of the 24 studies, 13 (54%) were "nonsignificant" at $\alpha = 0.05$, one-tailed. But if the 367 trials in these "failed replications" are combined, there are 106 direct hits, $z = 1.66$, and $p = 0.0485$, one tailed. This is reminiscent of the dilemma of Professor B in Section 3.

Power is typically very low for these studies. The median sample size for the studies reporting direct hits was 28. If there is a real effect and it increases the success probability from the chance 0.25 to an actual 0.33 (a value whose rationale will be made clear below), the power for a study with 28 trials is only 0.181 (Utts, 1986). It should be no surprise that there is a "repeatability" problem in parapsychology.

4.4 Flaw Analysis and Future Recommendations

The second half of Hyman's paper consisted of a "Meta-Analysis of Flaws and Successful Outcomes" (1985b, page 30), designed to explore whether or not various measures of success were related to specific flaws in the experiments. While many critics have argued that the results in parapsychology can be explained by experimental flaws, Hyman's analysis was the first to attempt to quantify the relationship between flaws and significant results.

Hyman identified 12 potential flaws in the ganzfeld experiments, such as inadequate random-

ization, multiple tests used without adjusting the significance level (thus inflating the significance level from the nominal 5%) and failure to use a duplicate set of targets for the judging process (thus allowing possible clues such as fingerprints). Using cluster and factor analyses, the 12 binary flaw variables were combined into three new variables, which Hyman named General Security, Statistics and Controls.

Several analyses were then conducted. The one reported with the most detail is a factor analysis utilizing 17 variables for each of 36 studies. Four factors emerged from the analysis. From these, Hyman concluded that security had increased over the years, that the significance level tended to be inflated the most for the most complex studies and that both effect size and level of significance were correlated with the existence of flaws.

Following his factor analysis, Hyman picked the three flaws that seemed to be most highly correlated with success, which were inadequate attention to both randomization and documentation and the potential for ordinary communication between the sender and receiver. A regression equation was then computed using each of the three flaws as dummy variables, and the effect size for the experiment as the dependent variable. From this equation, Hyman concluded that a study without these three flaws would be predicted to have a hit rate of 27%. He concluded that this is "well within the statistical neighborhood of the 25% chance rate" (1985b, page 37), and thus "the ganzfeld psi data base, despite initial impressions, is inadequate either to support the contention of a repeatable study or to demonstrate the reality of psi" (page 38).

Honorton discounted both Hyman's flaw classification and his analysis. He did not deny that flaws existed, but he objected that Hyman's analysis was faulty and impossible to interpret. Honorton asked psychometrician David Saunders to write an Appendix to his article, evaluating Hyman's analysis. Saunders first criticized Hyman's use of a factor analysis with 17 variables (many of which were dichotomous) and only 36 cases and concluded that "the entire analysis is meaningless" (Saunders, 1985, page 87). He then noted that Hyman's choice of the three flaws to include in his regression analysis constituted a clear case of multiple analysis, since there were 84 possible sets of three that could have been selected (out of nine potential flaws), and Hyman chose the set most highly correlated with effect size. Again, Saunders concluded that "any interpretation drawn from [the regression analysis] must be regarded as meaningless" (1985, page 88).

Hyman's results were also contradicted by Harris and Rosenthal (1988b) in an analysis requested by Hyman in his capacity as Chair of the National Academy of Sciences' Subcommittee on Parapsychology. Using Hyman's flaw classifications and a multivariate analysis, Harris and Rosenthal concluded that "Our analysis of the effects of flaws on study outcome lends no support to the hypothesis that ganzfeld research results are a significant function of the set of flaw variables" (1988b, page 3).

Hyman and Honorton were in the process of preparing papers for a second round of debate when they were invited to lunch together at the 1986 Meeting of the Parapsychological Association. They discovered that they were in general agreement on several major issues, and they decided to coauthor a "Joint Communiqué" (Hyman and Honorton, 1986). It is clear from their paper that they both thought it was more important to set the stage for future experimentation than to continue the technical arguments over the current data base. In the abstract to their paper, they wrote:

> We agree that there is an overall significant effect in this data base that cannot reasonably be explained by selective reporting or multiple analysis. We continue to differ over the degree to which the effect constitutes evidence for psi, but we agree that the final verdict awaits the outcome of future experiments conducted by a broader range of investigators and according to more stringent standards [page 351].

The paper then outlined what these standards should be. They included controls against any kind of sensory leakage, thorough testing and documentation of randomization methods used, better reporting of judging and feedback protocols, control for multiple analyses and advance specification of number of trials and type of experiment. Indeed, any area of research could benefit from such a careful list of procedural recommendations.

4.5 Rosenthal's Meta-Analysis

The same issue of the *Journal of Parapsychology* in which the Joint Communiqué appeared also carried commentaries on the debate by 10 separate authors. In his commentary, psychologist Robert Rosenthal, one of the pioneers of meta-analysis in psychology, summarized the aspects of Hyman's and Honorton's work that would typically be included in a meta-analysis (Rosenthal, 1986). It is worth reviewing Rosenthal's results so that they can be used as a basis of comparison for the more recent psi ganzfeld studies reported in Section 5.

Rosenthal, like Hyman and Honorton, focused only on the 28 studies for which direct hits were known. He chose to use an effect size measure

called Cohen's h, which is the difference between the arcsin transformed proportions of direct hits that were observed and expected:

$$h = 2(\arcsin \sqrt{\hat{p}} - \arcsin \sqrt{p}).$$

One advantage of this measure over the difference in raw proportions is that it can be used to compare experiments with different chance hit rates.

If the observed and expected numbers of hits were identical, the effect size would be zero. Of the 28 studies, 23 (82%) had effect sizes greater than zero, with a median effect size of 0.32 and a mean of 0.28. These correspond to direct hit rates of 0.40 and 0.38 respectively, when 0.25 is expected by chance. A 95% confidence interval for the true effect size is from 0.11 to 0.45, corresponding to direct hit rates of from 0.30 to 0.46 when chance is 0.25.

A common technique in meta-analysis is to calculate a "combined z," found by summing the individual z scores and dividing by the square root of the number of studies. The result should have a standard normal distribution if each z score has a standard normal distribution. For the ganzfeld studies, Rosenthal reported a combined z of 6.60 with a p-value of 3.37×10^{-11}. He also reiterated Honorton's file-drawer assessment by calculating that there would have to be 423 studies unreported to negate the significant effect in the 28 direct hit studies.

Finally, Rosenthal acknowledged that, because of the flaws in the data base and the potential for at least a small file-drawer effect, the true average effect size was probably closer to 0.18 than 0.28. He concluded, "Thus, when the accuracy rate expected under the null is 1/4, we might estimate the obtained accuracy rate to be about 1/3" (1986, page 333). This is the value used for the earlier power calculation.

It is worth mentioning that Rosenthal was commissioned by the National Academy of Sciences to prepare a background paper to accompany its 1988 report on parapsychology. That paper (Harris and Rosenthal, 1988a) contained much of the same analysis as his commentary summarized above. Ironically, the discussion of the ganzfeld work in the National Academy Report focused on Hyman's 1985 analysis, but never mentioned the work it had commissioned Rosenthal to perform, which contradicted the final conclusion in the report.

5. A META-ANALYSIS OF RECENT GANZFELD EXPERIMENTS

After the initial exchange with Hyman at the 1982 Parapsychological Association Meeting, Honorton and his colleagues developed an automated ganzfeld experiment that was designed to eliminate the methodological flaws identified by Hyman. The execution and reporting of the experiments followed the detailed guidelines agreed upon by Hyman and Honorton.

Using this "autoganzfeld" experiment, 11 experimental series were conducted by eight experimenters between February 1983 and September 1989, when the equipment had to be dismantled due to lack of funding. In this section, the results of these experiments are summarized and compared to the earlier ganzfeld studies. Much of the information is derived from Honorton et al. (1990).

5.1 The Automated Ganzfeld Procedure

Like earlier ganzfeld studies, the "autoganzfeld" experiments require four participants. The first is the Receiver (R), who attempts to identify the target material being observed by the Sender (S). The Experimenter (E) prepares R for the task, elicits the response from R and supervises R's judging of the response against the four potential targets. (Judging is double blind; E does not know which is the correct target.) The fourth participant is the lab assistant (LA) whose only task is to instruct the computer to randomly select the target. No one involved in the experiment knows the identity of the target.

Both R and S are sequestered in sound-isolated, electrically shielded rooms. R is prepared as in earlier ganzfeld studies, with white noise and a field of red light. In a nonadjacent room, S watches the target material on a television and can hear R's target description ("mentation") as it is being given. The mentation is also tape recorded.

The judging process takes place immediately after the 30-minute sending period. On a TV monitor in the isolated room, R views the four choices from the target pack that contains the actual target. R is asked to rate each one according to how closely it matches the ganzfeld mentation. The ratings are converted to ranks and, if the correct target is ranked first, a direct hit is scored. The entire process is automatically recorded by the computer. The computer then displays the correct choice to R as feedback.

There were 160 preselected targets, used with replacement, in 10 of the 11 series. They were arranged in packets of four, and the decoys for a given target were always the remaining three in the same set. Thus, even if a particular target in a set were consistently favored by Rs, the probability of a direct hit under the null hypothesis would remain at 1/4. Popular targets should be no more

likely to be selected by the computer's random number generator than any of the others in the set. The selection of the target by the computer is the only source of randomness in these experiments. This is an important point, and one that is often misunderstood. (See Utts, 1991, for elucidation.)

Eighty of the targets were "dynamic," consisting of scenes from movies, documentaries and cartoons; 80 were "static," consisting of photographs, art prints and advertisements. The four targets within each set were all of the same type. Earlier studies indicated that dynamic targets were more likely to produce successful results, and one of the goals of the new experiments was to test that theory.

The randomization procedure used to select the target and the order of presentation for judging was thoroughly tested before and during the experiments. A detailed description is given by Honorton et al. (1990, pages 118–120).

Three of the 11 series were pilot series, five were formal series with novice receivers, and three were formal series with experienced receivers. The last series with experienced receivers was the only one that did not use the 160 targets. Instead, it used only one set of four dynamic targets in which one target had previously received several first place ranks and one had never received a first place rank. The receivers, none of whom had had prior exposure to that target pack, were not aware that only one target pack was being used. They each contributed one session only to the series. This will be called the "special series" in what follows.

Except for two of the pilot series, numbers of trials were planned in advance for each series. Unfortunately, three of the formal series were not yet completed when the funding ran out, including the special series, and one pilot study with advance planning was terminated early when the experimenter relocated. There were no unreported trials during the 6-year period under review, so there was no "file drawer."

Overall, there were 183 Rs who contributed only one trial and 58 who contributed more than one, for a total of 241 participants and 355 trials. Only 23 Rs had previously participated in ganzfeld experiments, and 194 Rs (81%) had never participated in any parapsychological research.

5.2 Results

While acknowledging that no probabilistic conclusions can be drawn from qualitative data, Honorton et al. (1990) included several examples of session excerpts that Rs identified as providing the basis for their target rating. To give a flavor for the dream-like quality of the mentation and the amount of information that can be lost by only assigning a rank, the first example is reproduced here. The target was a painting by Salvador Dali called "Christ Crucified." The correct target received a first place rank. The part of the mentation R used to make this assessment read:

> . . . I think of guides, like spirit guides, leading me and I come into a court with a king. It's quiet . . . . It's like heaven. The king is something like Jesus. Woman. Now I'm just sort of summersaulting through heaven . . . . Brooding . . . . Aztecs, the Sun God . . . . High priest . . . .Fear . . . . Graves. Woman. Prayer . . . . Funeral . . . . Dark. Death . . . . Souls . . . . Ten Commandments. Moses . . . . [Honorton et al., 1990].

Over all 11 series, there were 122 direct hits in the 355 trials, for a hit rate of 34.4% (exact binomial p-value = 0.00005) when 25% were expected by chance. Cohen's h is 0.20, and a 95% confidence interval for the overall hit rate is from 0.30 to 0.39. This calculation assumes, of course, that the probability of a direct hit is constant and independent across trials, an assumption that may be questionable except under the null hypothesis of no psi abilities.

Honorton et al. (1990) also calculated effect sizes for each of the 11 series and each of the eight experimenters. All but one of the series (the first novice series) had positive effect sizes, as did all of the experimenters.

The special series with experienced Rs had an exceptionally high effect size with $h = 0.81$, corresponding to 16 direct hits out of 25 trials (64%), but the remaining series and the experimenters had relatively homogeneous effect sizes given the amount of variability expected by chance. If the special series is removed, the overall hit rate is 32.1%, $h = 0.16$. Thus, the positive effects are not due to just one series or one experimenter.

Of the 218 trials contributed by novices, 71 were direct hits (32.5%, $h = 0.17$), compared with 51 hits in the 137 trials by those with prior ganzfeld experience (37%, $h = 0.26$). The hit rates and effect sizes were 31% ($h = 0.14$) for the combined pilot series, 32.5% ($h = 0.17$) for the combined formal novice series, and 41.5% ($h = 0.35$) for the combined experienced series. The last figure drops to 31.6% if the outlier series is removed. Finally, without the outlier series the hit rate for the combined series where all of the planned trials were completed was 31.2% ($h = 0.14$), while it was 35% ($h = 0.22$) for the combined series that were terminated early. Thus, optional stopping cannot account for the positive effect.

There were two interesting comparisons that had been suggested by earlier work and were preplanned in these experiments. The first was to compare results for trials with dynamic targets with those for static targets. In the 190 dynamic target sessions there were 77 direct hits (40%, $h = 0.32$) and for the static targets there were 45 hits in 165 trials (27%, $h = 0.05$), thus indicating that dynamic targets produced far more successful results.

The second comparison of interest was whether or not the sender was a friend of the receiver. This was a choice the receiver could make. If he or she did not bring a friend, a lab member acted as sender. There were 211 trials with friends as senders (some of whom were also lab staff), resulting in 76 direct hits (36%, $h = 0.24$). Four trials used no sender. The remaining 140 trials used nonfriend lab staff as senders and resulted in 46 direct hits (33%, $h = 0.18$). Thus, trials with friends as senders were slightly more successful than those without.

Consonant with the definition of replication based on consistent effect sizes, it is informative to compare the autoganzfeld experiments with the direct hit studies in the previous data base. The overall success rates are extremely similar. The overall direct hit rate was 34.4% for the autoganzfeld studies and was 38% for the comparable direct hit studies in the earlier meta-analysis. Rosenthal's (1986) adjustment for flaws had placed a more conservative estimate at 33%, very close to the observed 34.4% in the new studies.

One limitation of this work is that the autoganzfeld studies, while conducted by eight experimenters, all used the same equipment in the same laboratory. Unfortunately, the level of funding available in parapsychology and the cost in time and equipment to conduct proper experiments make it difficult to amass large amounts of data across laboratories. Another autoganzfeld laboratory is currently being constructed at the University of Edinburgh in Scotland, so interlaboratory comparisons may be possible in the near future.

Based on the effect size observed to date, large samples are needed to achieve reasonable power. If there is a constant effect across all trials, resulting in 33% direct hits when 25% are expected by chance, to achieve a one-tailed significance level of 0.05 with 95% probability would require 345 sessions.

We end this section by returning to the aspirin and heart attack example in Section 3 and expanding a comparison noted by Atkinson, Atkinson, Smith and Bem (1990, page 237). Computing the equivalent of Cohen's h for comparing observed heart attack rates in the aspirin and placebo groups results in $h = 0.068$. Thus, the effect size observed in the ganzfeld data base is triple the much publicized effect of aspirin on heart attacks.

6. OTHER META-ANALYSES IN PARAPSYCHOLOGY

Four additional meta-analyses have been conducted in various areas of parapsychology since the original ganzfeld meta-analyses were reported. Three of the four analyses focused on evidence of psi abilities, while the fourth examined the relationship between extroversion and psychic functioning. In this section, each of the four analyses will be briefly summarized.

There are only a handful of English-language journals and proceedings in parapsychology, so retrieval of the relevant studies in each of the four cases was simple to accomplish by searching those sources in detail and by searching other bibliographic data bases for keywords.

Each analysis included an overall summary, an analysis of the quality of the studies versus the size of the effect and a "file-drawer" analysis to determine the possible number of unreported studies. Three of the four also contained comparisons across various conditions.

6.1 Forced-Choice Precognition Experiments

Honorton and Ferrari (1989) analyzed forced-choice experiments conducted from 1935 to 1987, in which the target material was randomly selected *after* the subject had attempted to predict what it would be. The time delay in selecting the target ranged from under a second to one year. Target material included items as diverse as ESP cards and automated random number generators. Two investigators, S. G. Soal and Walter J. Levy, were not included because some of their work has been suspected to be fraudulent.

Overall Results. There were 309 studies reported by 62 senior authors, including more than 50,000 subjects and nearly two million individual trials. Honorton and Ferrari used $z/\sqrt{n}$ as the measure of effect size (ES) for each study, where n was the number of Bernoulli trials in the study. They reported a mean ES of 0.020, and a mean z-score of 0.65 over all studies. They also reported a combined z of 11.41, $p = 6.3 \times 10^{-25}$. Some 30% (92) of the studies were statistically significant at $\alpha = 0.05$. The mean ES per investigator was 0.033, and the significant results were not due to just a few investigators.

Quality. Eight dichotomous quality measures were assigned to each study, resulting in possible

scores from zero for the lowest quality, to eight for the highest. They included features such as adequate randomization, preplanned analysis and automated recording of the results. The correlation between study quality and effect size was 0.081, indicating a slight tendency for higher quality studies to be more successful, contrary to claims by critics that the opposite would be true. There was a clear relationship between quality and year of publication, presumably because over the years experimenters in parapsychology have responded to suggestions from critics for improving their methodology.

File Drawer. Following Rosenthal (1984), the authors calculated the "fail-safe N" indicating the number of unreported studies that would have to be sitting in file drawers in order to negate the significant effect. They found $N = 14{,}268$, or a ratio of 46 unreported studies for each one reported. They also followed a suggestion by Dawes, Landman and Williams (1984) and computed the mean z for all studies with $z > 1.65$. If such studies were a random sample from the upper 5% tail of a $N(0, 1)$ distribution, the mean z would be 2.06. In this case it was 3.61. They concluded that selective reporting could not explain these results.

Comparisons. Four variables were identified that appeared to have a systematic relationship to study outcome. The first was that the 25 studies using subjects selected on the basis of good past performance were more successful than the 223 using unselected subjects, with mean effect sizes of 0.051 and 0.008, respectively. Second, the 97 studies testing subjects individually were more successful than the 105 studies that used group testing; mean effect sizes were 0.021 and 0.004, respectively. Timing of feedback was the third moderating variable, but information was only available for 104 studies. The 15 studies that never told the subjects what the targets were had a mean effect size of −0.001. Feedback after each trial produced the best results, the mean ES for the 47 studies was 0.035. Feedback after each set of trials resulted in mean ES of 0.023 (21 studies), while delayed feedback (also 21 studies) yielded a mean ES of only 0.009. There is a clear ordering; as the gap between time of feedback and time of the actual guesses decreased, effect sizes increased.

The fourth variable was the time interval between the subject's guess and the actual target selection, available for 144 studies. The best results were for the 31 studies that generated targets less than a second after the guess (mean $ES = 0.045$), while the worst were for the seven studies that delayed target selection by at least a month (mean $ES = 0.001$). The mean effect sizes showed a clear trend, decreasing in order as the time interval increased from minutes to hours to days to weeks to months.

6.2 Attempts to Influence Random Physical Systems

Radin and Nelson (1989) examined studies designed to test the hypothesis that "The statistical output of an electronic RNG [random number generator] is correlated with observer intention in accordance with prespecified instructions" (page 1502). These experiments typically involve RNGs based on radioactive decay, electronic noise or pseudorandom number sequences seeded with true random sources. Usually the subject is instructed to try to influence the results of a string of binary trials by mental intention alone. A typical protocol would ask a subject to press a button (thus starting the collection of a fixed-length sequence of bits), and then try to influence the random source to produce more zeroes or more ones. A run might consist of three successive button presses, one each in which the desired result was more zeroes or more ones, and one as a control with no conscious intention. A z score would then be computed for each button press.

The 832 studies in the analysis were conducted from 1959 to 1987 and included 235 "control" studies, in which the output of the RNGs were recorded but there was no conscious intention involved. These were usually conducted before and during the experimental series, as tests of the RNGs.

Results. The effect size measure used was again $z/\sqrt{n}$, where z was positive if more bits of the specified type were achieved. The mean effect size for control studies was not significantly different from zero (-1.0×10^{-5}). The mean effect size for the experimental studies was also very small, 3.2×10^{-4}, but it was significantly higher than the mean ES for the control studies ($z = 4.1$).

Quality. Sixteen quality measures were defined and assigned to each study, under the four general categories of procedures, statistics, data and the RNG device. A score of 16 reflected the highest quality. The authors regressed mean effect size on mean quality for each investigator and found a slope of 2.5×10^{-5} with standard error of 3.2×10^{-5}, indicating little relationship between quality and outcome. They also calculated a weighted mean effect size, using quality scores as weights, and found that it was very similar to the unweighted mean ES. They concluded that "differences in methodological quality are not significant predictors of effect size" (page 1507).

File Drawer. Radin and Nelson used several methods for estimating the number of unreported

studies (pages 1508–1510). Their estimates ranged from 200 to 1000 based on models assuming that all significant studies were reported. They calculated the fail-safe N to be 54,000.

6.3 Attempts to Influence Dice

Radin and Ferrari (1991) examined 148 studies, published from 1935 to 1987, designed to test whether or not consciousness can influence the results of tossing dice. They also found 31 "control" studies in which no conscious intention was involved.

Results. The effect size measure used was $z/\sqrt{n}$, where z was based on the number of throws in which the die landed with the desired face (or faces) up, in n throws. The weighted mean ES for the experimental studies was 0.0122 with a standard error of 0.00062; for the control studies the mean and standard error were 0.00093 and 0.00255, respectively. Weights for each study were determined by quality, giving more weight to high-quality studies. Combined z scores for the experimental and control studies were reported by Radin and Ferrari to be 18.2 and 0.18, respectively.

Quality. Eleven dichotomous quality measures were assigned, ranging from automated recording to whether or not control studies were interspersed with the experimental studies. The final quality score for each study combined these with information on method of tossing the dice, and with source of subject (defined below). A regression of quality score versus effect size resulted in a slope of -0.002, with a standard error of 0.0011. However, when effect sizes were weighted by sample size, there was a significant relationship between quality and effect size, leading Radin and Ferrari to conclude that higher-quality studies produced lower weighted effect sizes.

File Drawer. Radin and Ferrari calculated Rosenthal's fail-safe. N for this analysis to be 17,974. Using the assumption that all significant studies were reported, they estimated the number of unreported studies to be 1152. As a final assessment, they compared studies published before and after 1975, when the *Journal of Parapsychology* adopted an official policy of publishing nonsignificant results. They concluded, based on that analysis, that more nonsignificant studies were published after 1975, and thus "We must consider the overall (1935–1987) data base as suspect with respect to the filedrawer problem."

Comparisons. Radin and Ferrari noted that there was bias in both the experimental and control studies across die face. Six was the face most likely to come up, consistent with the observation that it has the least mass. Therefore, they examined results for the subset of 69 studies in which targets were evenly balanced among the six faces. They still found a significant effect, with mean and standard error for effect size of 8.6×10^{-3} and 1.1×10^{-3}, respectively. The combined z was 7.617 for these studies.

They also compared effect sizes across types of subjects used in the studies, categorizing them as unselected, experimenter and other subjects, experimenter as sole subject, and specially selected subjects. Like Honorton and Ferrari (1989), they found the highest mean ES for studies with selected subjects; it was approximately 0.02, more than twice that for unselected subjects.

6.4 Extroversion and ESP Performance

Honorton, Ferrari and Bem (1991) conducted a meta-analysis to examine the relationship between scores on tests of extroversion and scores on psi-related tasks. They found 60 studies by 17 investigators, conducted from 1945 to 1983.

Results. The effect size measure used for this analysis was the correlation between each subject's extroversion score and ESP score. A variety of measures had been used for both scores across studies, so various correlation coefficients were used. Nonetheless, a stem and leaf diagram of the correlations showed an approximate bell shape with mean and standard deviation of 0.19 and 0.26, respectively, and with an additional outlier at $r = 0.91$. Honorton et al. reported that when weighted by degrees of freedom, the weighted mean r was 0.14, with a 95% confidence interval covering 0.10 to 0.19.

Forced-Choice versus Free-Response Results. Because forced-choice and free-response tests differ qualitatively, Honorton et al. chose to examine their relationship to extroversion separately. They found that for free-response studies there was a significant correlation between extroversion and ESP scores, with mean $r = 0.20$ and $z = 4.46$. Further, this effect was homogeneous across both investigators and extroversion scales.

For forced-choice studies, there was a significant correlation between ESP and extroversion, but only for those studies that reported the ESP results to the subjects *before* measuring extroversion. Honorton et al. speculated that the relationship was an artifact, in which extroversion scores were temporarily inflated as a result of positive feedback on ESP performance.

Confirmation with New Data Following the extroversion/ESP meta-analysis, Honorton et al. attempted to confirm the relationship using the autoganzfeld data base. Extroversion scores based on the Myers–Briggs Type Indicator were available for 221 of the 241 subjects who had participated in autoganzfeld studies.

The correlation between extroversion scores and ganzfeld rating scores was $r = 0.18$, with a 95% confidence interval from 0.05 to 0.30. This is consistent with the mean correlation of $r = 0.20$ for free-response experiments, determined from the meta-analysis. These correlations indicate that extroverted subjects can produce higher scores in free-response ESP tests.

7. CONCLUSIONS

Parapsychologists often make a distinction between "proof-oriented research" and "process-oriented research." The former is typically conducted to test the hypothesis that psi abilities exist, while the latter is designed to answer questions about how psychic functioning works. Proof-oriented research has dominated the literature in parapsychology. Unfortunately, many of the studies used small samples and would thus be nonsignificant even if a moderate-sized effect exists.

The recent focus on meta-analysis in parapsychology has revealed that there are small but consistently nonzero effects across studies, experimenters and laboratories. The sizes of the effects in forced-choice studies appear to be comparable to those reported in some medical studies that had been heralded as breakthroughs. (See Section 5; also Honorton and Ferrari, 1989, page 301.) Free-response studies show effect sizes of far greater magnitude.

A promising direction for future process-oriented research is to examine the causes of individual differences in psychic functioning. The ESP/extroversion meta-analysis is a step in that direction.

In keeping with the idea of individual differences, Bayes and empirical Bayes methods would appear to make more sense than the classical inference methods commonly used, since they would allow individual abilities and beliefs to be modeled. Jeffreys (1990) reported a Bayesian analysis of some of the RNG experiments and showed that conclusions were closely tied to prior beliefs even though hundreds of thousands of trials were available.

It may be that the nonzero effects observed in the meta-analyses can be explained by something other than ESP, such as shortcomings in our understanding of randomness and independence. Nonetheless, there is an anomaly that needs an explanation. As I have argued elsewhere (Utts, 1987), research in parapsychology should receive more support from the scientific community. If ESP does not exist, there is little to be lost by erring in the direction of further research, which may in fact uncover other anomalies. If ESP does exist, there is much to be lost by not doing process-oriented research, and much to be gained by discovering how to enhance and apply these abilities to important world problems.

ACKNOWLEDGMENTS

I would like to thank Deborah Delanoy, Charles Honorton, Wesley Johnson, Scott Plous and an anonymous reviewer for their helpful comments on an earlier draft of this paper, and Robert Rosenthal and Charles Honorton for discussions that helped clarify details.

REFERENCES

ATKINSON, R. L., ATKINSON, R. C., SMITH, E. E. and BEM, D. J. (1990). *Introduction to Psychology*, 10th ed. Harcourt Brace Jovanovich, San Diego.

BELOFF, J. (1985). Research strategies for dealing with unstable phenomena. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 1–21. Parapsychology Foundation, New York.

BLACKMORE, S. J. (1985). Unrepeatability: Parapsychology's only finding. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 183–206. Parapsychology Foundation, New York.

BURDICK, D. S. and KELLY, E. F. (1977). Statistical methods in parapsychological research. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 81–130. Van Nostrand Reinhold, New York.

CAMP, B. H. (1937). (Statement in Notes Section.) *Journal of Parapsychology* **1** 305.

COHEN, J. (1990). Things I have learned (so far). *American Psychologist* **45** 1304–1312.

COOVER, J. E. (1917). *Experiments in Psychical Research at Leland Stanford Junior University*. Stanford Univ.

DAWES, R. M., LANDMAN, J. and WILLIAMS, J. (1984). Reply to Kurosawa. *American Psychologist* **39** 74–75.

DIACONIS, P. (1978). Statistical problems in ESP research. *Science* **201** 131–136.

DOMMEYER, F. C. (1975). Psychical research at Stanford University. *Journal of Parapsychology* **39** 173–205.

DRUCKMAN, D. and SWETS, J. A., eds. (1988) *Enhancing Human Performance: Issues, Theories, and Techniques*. National Academy Press, Washington, D.C.

EDGEWORTH, F. Y. (1885). The calculus of probabilities applied to psychical research. In *Proceedings of the Society for Psychical Research* **3** 190–199.

EDGEWORTH, F. Y. (1886). The calculus of probabilities applied to psychical research. II. In *Proceedings of the Society for Psychical Research* **4** 189–208.

FELLER, W. K. (1940). Statistical aspects of ESP. *Journal of Parapsychology* **4** 271–297.

FELLER, W. K. (1968). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York.

FISHER, R. A. (1924). A method of scoring coincidences in tests with playing cards. In *Proceedings of the Society for Psychical Research* **34** 181–185.

FISHER, R. A. (1929). The statistical method in psychical research. In *Proceedings of the Society for Psychical Research* **39** 189–192.

GALLUP, G. H., JR., and NEWPORT, F. (1991). Belief in paranormal phenomena among adult Americans. *Skeptical Inquirer* **15** 137–146.

GARDNER, M. J. and ALTMAN, D. G. (1986). Confidence intervals rather than *p*-values: Estimation rather than hypothesis testing. *British Medical Journal* **292** 746–750.

GILMORE, J. B. (1989). Randomness and the search for psi. *Journal of Parapsychology* **53** 309–340.

GILMORE, J. B. (1990). Anomalous significance in pararandom and psi-free domains. *Journal of Parapsychology* **54** 53–58.

GREELEY, A. (1987). Mysticism goes mainstream. *American Health* **7** 47–49.

GREENHOUSE, J. B. and GREENHOUSE, S. W. (1988). An aspirin a day . . . ? *Chance* **1** 24–31.

GREENWOOD, J. A. and STUART, C. E. (1940). A review of Dr. Feller's critique. *Journal of Parapsychology* **4** 299–319.

HACKING, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis* **79** 427–451.

HANSEL, C. E. M. (1980). *ESP and Parapsychology: A Critical Re-evaluation*. Prometheus Books, Buffalo, N.Y.

HARRIS, M. J. and ROSENTHAL, R. (1988a). *Interpersonal Expectancy Effects and Human Performance Research*. National Academy Press, Washington, D.C.

HARRIS, M. J. and ROSENTHAL, R. (1988b). *Postscript to Interpersonal Expectancy Effects and Human Performance Research*. National Academy Press, Washington, D.C.

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic, Orlando, Fla.

HONORTON, C. (1977). Psi and internal attention states. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 435–472. Van Nostrand Reinhold, New York.

HONORTON, C. (1985a). How to evaluate and improve the replicability of parapsychological effects. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 238–255. Parapsychology Foundation, New York.

HONORTON, C. (1985b). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology* **49** 51–91.

HONORTON, C., BERGER, R. E., VARVOGLIS, M. P., QUANT, M., DERR, P., SCHECHTER, E. I. and FERRARI, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology* **54** 99–139.

HONORTON, C. and FERRARI, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935–1987. *Journal of Parapsychology* **53** 281–308.

HONORTON, C., FERRARI, D. C. and BEM, D. J. (1991). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Research in Parapsychology 1990*. The Scarecrow Press, Metuchen, N.J. To appear.

HYMAN, R. (1985a). A critical overview of parapsychology. In *A Skeptic's Handbook of Parapsychology* (P. Kurtz, ed.) 1–96. Prometheus Books, Buffalo, N.Y.

HYMAN, R. (1985b). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology* **49** 3–49.

HYMAN, R. and HONORTON, C. (1986). Joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology* **50** 351–364.

IVERSEN, G. R., LONGCOR, W. H., MOSTELLER, F., GILBERT, J. P. and YOUTZ, C. (1971). Bias and runs in dice throwing and recording: A few million throws. *Psychometrika* **36** 1–19.

JEFFREYS, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration* **4** 153–169.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192.

MAUSKOPF, S. H. and MCVAUGH, M. (1979). *The Elusive Science: Origins of Experimental Psychical Research*. Johns Hopkins Univ. Press.

MCVAUGH, M. R. and MAUSKOPF, S. H. (1976). J. B. Rhine's *Extrasensory Perception* and its background in psychical research. *Isis* **67** 161–189.

NEULIEP, J. W., ed. (1990). Handbook of replication research in the behavioral and social sciences. *Journal of Social Behavior and Personality* **5** (4) 1–510.

OFFICE OF TECHNOLOGY ASSESSMENT (1989). Report of a workshop on experimental parapsychology. *Journal of the American Society for Psychical Research* **83** 317–339.

PALMER, J. (1989). A reply to Gilmore. *Journal of Parapsychology* **53** 341–344.

PALMER, J. (1990). Reply to Gilmore: Round two. *Journal of Parapsychology* **54** 59–61.

PALMER, J. A., HONORTON, C. and UTTS, J. (1989). Reply to the National Research Council study on parapsychology. *Journal of the American Society for Psychical Research* **83** 31–49.

RADIN, D. I. and FERRARI, D. C. (1991). Effects of consciousness on the fall of dice: A meta-analysis. *Journal of Scientific Exploration* **5** 61–83.

RADIN, D. I. and NELSON, R. D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics* **19** 1499–1514.

RAO, K. R. (1985). Replication in conventional and controversial sciences. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 22–41. Parapsychology Foundation, New York.

RHINE, J. B. (1934). *Extrasensory Perception*. Boston Society for Psychical Research, Boston. (Reprinted by Branden Press, 1964.)

RHINE, J. B. (1977). History of experimental studies. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 25–47. Van Nostrand Reinhold, New York.

RICHET, C. (1884). La suggestion mentale et le calcul des probabilités. *Revue Philosophique* **18** 608–674.

ROSENTHAL, R. (1984). *Meta-Analytic Procedures for Social Research*. Sage, Beverly Hills.

ROSENTHAL, R. (1986). Meta-analytic procedures and the nature of replication: The ganzfeld debate. *Journal of Parapsychology* **50** 315–336.

ROSENTHAL, R. (1990a). How are we doing in soft psychology? *American Psychologist* **45** 775–777.

ROSENTHAL, R. (1990b). Replication in behavioral research. *Journal of Social Behavior and Personality* **5** 1–30.

SAUNDERS, D. R. (1985). On Hyman's factor analysis. *Journal of Parapsychology* **49** 86–88.

SHAPIN, B. and COLY, L., eds. (1985). *The Repeatability Problem in Parapsychology*. Parapsychology Foundation, New York.

SPENCER-BROWN, G. (1957). *Probability and Scientific Inference*. Longmans Green, London and New York.

STUART, C. E. and GREENWOOD, J. A. (1937). A review of criticisms of the mathematical evaluation of ESP data. *Journal of Parapsychology* **1** 295–304.

TVERSKY, A. and KAHNEMAN, D. (1982). Belief in the law of small numbers. In *Judgment Under Uncertainty: Heuristics and Biases* (D. Kahneman, P. Slovic and A. Tversky, eds.) 23–31. Cambridge Univ. Press.

UTTS, J. (1986). The ganzfeld debate: A statistician's perspective. *Journal of Parapsychology* **50** 395–402.

UTTS, J. (1987). Psi, statistics, and society. *Behavioral and Brain Sciences* **10** 615–616.

UTTS, J. (1988). Successful replication versus statistical significance. *Journal of Parapsychology* **52** 305–320.

UTTS, J. (1989). Randomness and randomization tests: A reply to Gilmore. *Journal of Parapsychology* **53** 345–351.

UTTS, J. (1991). Analyzing free-response data: A progress report. In *Psi Research Methodology: A Re-examination* (L. Coly, ed.). Parapsychology Foundation, New York. To appear.

WILKS, S. S. (1965a). Statistical aspects of expeirments in telepath. *N.Y. Statistician* **16** (6) 1–3.

WILKS, S. S. (1965b). Statistical aspects of experiments in telepathy. *N.Y. Statistician* **16** (7) 4–6.